CHAMELEON meets spatial cognition

Paul Mc Kevitt*
Center for PersonKommunikation (CPK)
Institute for Electronic Systems (IES)
Fredrik Bajers Vej 7A-6, Aalborg University
DK-9220, Aalborg, DENMARK.
E-mail: pmck@cpk.auc.dk

Abstract

Intelligent MultiMedia (IntelliMedia) focusses on the computer processing and understanding of input from at least speech, text and visual images in terms of semantic representations. We have developed a general suite of tools in the form of a software and hardware platform called *CHAMELEON* that can be tailored to conducting IntelliMedia in various application domains. CHAMELEON has an open distributed processing architecture and currently includes ten agent modules: blackboard, dialogue manager, domain model, gesture recogniser, laser system, microphone array, speech recogniser, speech synthesiser, natural language processor, and a distributed Topsy learner. Modules can communicate with each other and the blackboard which keeps a record of interactions over time via semantic representations in frames. Inputs to CHAMELEON can include synchronised spoken dialogue and images and outputs include synchronised laser pointing and spoken dialogue. An initial prototype application of CHAMELEON is an *IntelliMedia WorkBench* where a user can ask for information about things (e.g. 2D/3D models, pictures, objects, gadgets, people, or whatever) on a physical table. The current domain is a *Campus Information System* for 2D building plans which provides information about tenants, rooms and routes and can answer questions like "Whose office is this?" and "Show me the route from Paul Mc Kevitt's office to Paul Dalsgaard's office." in real time. Projective spatial relations are expected to occur often with the IntelliMedia WorkBench and Campus Information System and we give here a worked example of how the query "Who's in the office beside him?" is processed by the frame semantics, showing all frames appearing on the blackboard. CHAMELEON and the IntelliMedia WorkBench are ideal for testing integrated signal and symbol processing of spatial cognition, language and vision for the future of SuperinformationhighwayS.

1 Introduction

IntelliMedia, which involves the computer processing and understanding of perceptual input from at least speech, text and visual images, and then reacting to it, is complex and involves signal and symbol processing techniques from not just engineering and computer science but

*Paul Mc Kevitt is also a British Engineering and Physical Sciences Research Council (EPSRC) Advanced Fellow at the University of Sheffield, England for five years under grant B/94/AF/1833 for the Integration of Natural Language, Speech and Vision Processing. This paper was completed whilst Paul Mc Kevitt was a Visiting Professor at LIMSI-CNRS, Orsay, France.

also artificial intelligence and cognitive science (Mc Kevitt 1994, 1995/1996, 1997a). With IntelliMedia systems, people can interact in spoken dialogues with machines, querying about what is being presented and even their gestures and body language can be interpreted.

People are able to combine the processing of language and vision with apparent ease. In particular, people can use words to describe a picture, and can reproduce a picture from a language description. Moreover, people can exhibit this kind of behaviour over a very wide range of input pictures and language descriptions. Although there are theories of how we process vision and language, there are few theories about how such processing is integrated. There have been large debates in Psychology and Philosophy with respect to the degree to which people store knowledge as propositions or pictures (Kosslyn and Pomerantz 1977, Pylyshyn 1973). Other recent moves towards integration are reported in Denis and Carfantan (1993), Mc Kevitt (1994, 1995/96) and Pentland (1993). It is often the case that when people use language about the visual environment they often need to refer to spatial relationships and they use prepositions to do so (Retz-Schmidt 1988, Zelinsky-Wibbelt 1993). Spatial relations are a central issue in the integration of natural language and vision processing (Maaβ 1996, Olivier 1995, 1996, 1997).

The Institute for Electronic Systems at Aalborg University, Denmark has expertise in the area of IntelliMedia and has already established an initiative on Multimodal and Multimedia User Interfaces (MMUI) called IntelliMedia 2000+ by the Faculty of Science and Technology. IntelliMedia 2000+ coordinates research on the production of a number of real-time demonstrators exhibiting examples of IntelliMedia applications, established a new Master's degree in IntelliMedia, and coordinates a nation-wide MultiMedia Network (MMN) concerned with technology transfer to industry. IntelliMedia 2000+ involves three departments and is coordinated from the Center for PersonKommunikation (CPK) which has a wealth of experience and expertise in spoken language processing, one of the central components of IntelliMedia, but also radio communications which would be useful for mobile applications. More details on IntelliMedia 2000+ can be found on WWW: http://www.cpk.auc.dk/CPK/MMUI/.

2 CHAMELEON and the IntelliMedia WorkBench

IntelliMedia 2000+ has developed the first prototype of an IntelliMedia software and hardware platform called CHAMELEON which is general enough to be used for a number of different applications. CHAMELEON demonstrates that existing software modules for (1) distributed processing and learning, (2) decision taking, (3) image processing, and (4) spoken dialogue processing can be interfaced to a single platform and act as communicating agent modules within it. CHAMELEON is independent of any particular application domain and the various modules can be distributed over different machines. Most of the modules are programmed in C++ and C.

2.1 IntelliMedia WorkBench

An initial application of CHAMELEON is the *IntelliMedia WorkBench* which is a hardware and software platform as shown in Figure 1.

One or more cameras and lasers can be mounted in the ceiling, microphone array placed on the wall and there is a table where things (objects, gadgets, people, pictures, 2D/3D models, building plans, or whatever) can be placed. The current domain is a *Campus Information System* which at present gives information on the architectural and functional layout of a

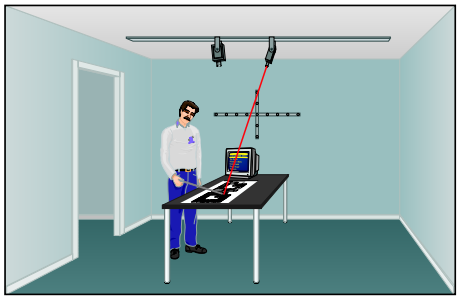

Figure 1: Physical layout of the IntelliMedia WorkBench

building. 2D architectural plans of the building drawn on white paper are laid on the table and the user can ask questions about them. At present the plans represent two floors of the 'A' (A2) building at Fredrik Bajers Vej 7, Aalborg University. The 2D plan is shown in Figure 2.

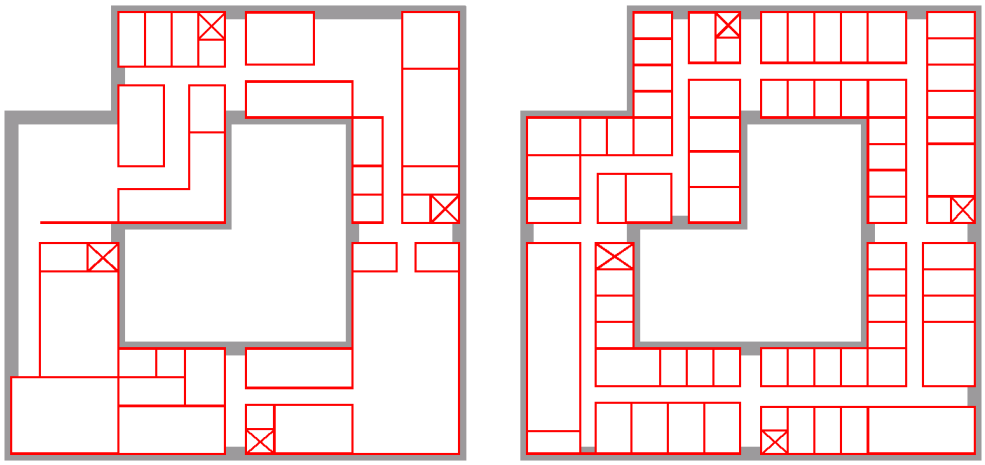

Figure 2: 2D plan of the 'A' building at Fredrik Bajers Vej 7, Aalborg University. Left: ground floor; Right: 1st floor.

Presently, there is one static camera which calibrates the plans on the table and the laser, and interprets the user's pointing while the system points to locations and draws routes with a laser. Inputs are simultaneous speech and/or pointing gestures and outputs are synchronised speech synthesis and pointing. We currently run all of CHAMELEON on a 200 MHz Intel pentium computer (r2d2) which handles input for the Campus Information System in real-time. More details on CHAMELEON and the IntelliMedia WorkBench can be found in Brøndsted et al. (1998).

2.2 Sample interaction dialogue

We present here a sample dialogue which the current first prototype can process. The example includes user intentions which are instructions and queries, and exophoric/deictic reference.

```
USER:      Show me Tom's office.
CHAMELEON: [points]
           This is Tom's office.
USER:      Point to Thomas' office.
CHAMELEON: [points]
           This is Thomas' office.
USER:      Where is the computer room?
CHAMELEON: [points]
           The computer room is here.
USER:      [points to instrument repair]
           Whose office is this?
CHAMELEON: [points]
           This is not an office, this is instrument repair.
USER:      [points]
           Whose office is this?
CHAMELEON: [points]
           This is Paul's office.
USER:      Show me the route from Lars Bo Larsen's office to Hanne Gade's office.
CHAMELEON: [draws route]
           This is the route from Lars Bo's office to Hanne's office.
USER:      Show me the route from Paul Mc Kevitt's office to instrument repair.
CHAMELEON: [draws route]
           This is the route from Paul's office to instrument repair.
USER:      Show me Paul's office.
CHAMELEON: [points]
           This is Paul's office.
```

Note that CHAMELEON assumes Paul Dalsgaard as default Paul[1] although there are two Pauls. A later prototype of the system should become active here and ask the user a question by first pointing out that there are two Pauls and then asking which does he/she mean. CHAMELEON can process deictic reference ("Whose office is *this*?") which is one of the most frequently occurring phenomena in IntelliMedia. However, spatial relations (e.g. "Who's in the office *beside* him?") are another phenomenon occurring regularly in language and vision integration which are not yet implemented in CHAMELEON. Also, there are other projective spatial relations such as "left", "right", "above", "below", and queries like "Who's in the office two up from him?" which occur regularly.

2.3 Architecture of CHAMELEON

CHAMELEON has a distributed architecture of communicating agent modules processing inputs and outputs from different modalities and each of which can be tailored to a number of application domains. The process synchronisation and intercommunication for CHAMELEON modules is performed using the DACS (Distributed Applications Communication System) Inter Process Communication (IPC) software (Fink et al. 1996) which enables CHAMELEON modules to be glued together and distributed across a number of servers. Presently, there are ten software modules in CHAMELEON: blackboard, dialogue manager, domain model, gesture recogniser, laser system, microphone array, speech recogniser, speech synthesiser, natural language processor (NLP), and Topsy as shown in Figure 3. The blackboard and dialogue

[1] This is because Paul Dalsgaard is more senior :).

manager form the kernel of CHAMELEON. We shall now give a brief description of each module.

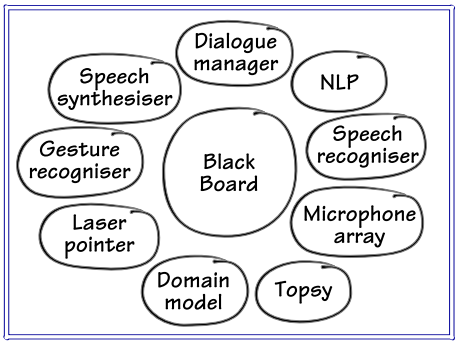

Figure 3: Architecture of CHAMELEON

The **blackboard** stores semantic representations produced by each of the other modules and keeps a history of these over the course of an interaction. All modules communicate through the exchange of semantic representations with each other or the blackboard. Semantic representations are frames in the spirit of Minsky (1975). The intention is that all modules in the system will produce and read frames. The frame semantics was first presented in Mc Kevitt and Dalsgaard (1997) and for the sample dialogue given in Section 2.2 CHAMELEON's actual blackboard history in terms of frames (messages) is shown in Appendix A.

The **dialogue manager** makes decisions about which actions to take and accordingly sends commands to the output modules (laser and speech synthesiser) via the blackboard. At present the functionality of the dialogue manager is to integrate and react to information coming in from the speech/NLP and gesture modules and to sending synchronised commands to the laser system and the speech synthesiser modules.

The **domain model** contains a database of all locations and their functionality, tenants and coordinates. The model is organised in a hierarchical structure: areas, buildings and rooms. Rooms are described by an identifier for the room (room number) and the type of the room (office, corridor, toilet, etc.). The model includes functions that return information about a room or a person. Possible inputs are coordinates or room number for rooms and name for persons, but in principle any attribute can be used as key and any other attribute can be returned. Furthermore, a path planner is provided, calculating the shortest route between two locations.

A design principle of imposing as few physical constraints as possible on the user (e.g. data gloves or touch screens) leads to the inclusion of a vision based **gesture recogniser**. Currently, it tracks a pointer via a camera mounted in the ceiling. Using one camera, the gesture recogniser is able to track 2D pointing gestures in real time. Only two gestures are recognised at present: pointing and not-pointing. From each digitised image the background is subtracted leaving only the motion (and some noise) within this image. This motion is analysed in order to find the direction of the pointing device and its tip. By temporal segmenting of these two parameters, a clear indication of the position the user is pointing to at a given time is found. The error of the tracker is less than one pixel (through an interpolation process) for the pointer.

A **laser system** acts as a "system pointer". It can be used for pointing to positions, drawing lines and displaying text. The laser beam is controlled in real-time (30 kHz). It can scan frames containing up to 600 points with a refresh rate of 50 Hz thus drawing very steady images on surfaces. It is controlled by a standard Pentium PC host computer. The pointer tracker and the laser pointer have been carefully calibrated so that they can work together. An automatic calibration procedure has been set up involving both the camera and laser where they are tested by asking the laser to follow the pointer.

A **microphone array** (Leth-Espensen and Lindberg 1996) is used to locate sound sources, e.g. a person speaking. Depending upon the placement of a maximum of 12 microphones it calculates sound source positions in 2D or 3D. It is based on measurement of the delays with which a sound wave arrives at the different microphones. From this information the location of the sound source can be identified. Another application of the array is to use it to focus at a specific location thus enhancing any acoustic activity at that location. This module is in the process of being incorporated into CHAMELEON.

Speech recognition is handled by the grapHvite real-time continuous speech recogniser (Power et al. 1997). It is based on HMMs (Hidden Markov Models) of triphones for acoustic decoding of English or Danish. The recognition process focusses on recognition of speech concepts and ignores non content words or phrases. A finite state network describing phrases is created by hand in accordance with the domain model and the grammar for the natural language parser. The latter can also be performed automatically by a grammar converter in the NLP module. The speech recogniser takes speech signals as input and produces text strings as output. Integration of the the latest CPK speech recogniser (Christensen et al. 1998) which is under development is being considered.

We use the Infovox Text-To-Speech (TTS) **speech synthesiser** which at present is capable of synthesising Danish and English (Infovox 1994). It is a rule based formant synthesiser and can simultaneously cope with multiple languages, e.g. pronounce a Danish name within an English utterance. Infovox takes text as input and produces speech as output. Integration of the the CPK speech synthesiser (Nielsen et al. 1997) which is under development for English is being considered.

Natural language processing is based on a compound feature based (so-called unification) grammar formalism for extracting semantics from the one-best utterance text output from the speech recogniser (Brøndsted 1998). The parser carries out a syntactic constituent analysis of input and subsequently maps values into semantic frames. The rules used for syntactic parsing are based on a subset of the EUROTRA formalism, i.e. in terms of lexical rules and structure building rules (Bech 1991). Semantic rules define certain syntactic subtrees and which frames to create if the subtrees are found in the syntactic parse trees. The natural language generator is currently under construction and at present generation is conducted by using canned text.

The basis of the Phase Web paradigm (Manthey 1998), and its incarnation in the form of a program called **Topsy**, is to represent knowledge and behaviour in the form of hierarchical relationships between the mutual exclusion and co-occurrence of events. In AI parlance, Topsy is a distributed, associative, continuous-action, dynamic partial-order planner that learns from experience. Relative to MultiMedia, integrating independent data from multiple media begins with noticing that what ties otherwise independent inputs together is the fact that they occur simultaneously (more or less). This is also Topsy's basic operating principle, but this is further combined with the notion of mutual exclusion, and thence to hierarchies of such relationships (Manthey 1998).

3 Frame semantics

The meaning of interactions over the course of a MultiModal dialogue is represented using a frame semantics with frames in the spirit of Minsky (1975). The intention is that all modules in the system can produce and read frames. Frames are coded in CHAMELEON with messages built as predicate-argument structures following a BNF definition. Frames represent some crucial elements such as *module*, *input/output*, *intention*, *location*, and *timestamp*. Module is simply the name of the module producing the frame (e.g. NLP). Inputs are the input recognised whether spoken (e.g. "Show me Hanne's office") or gestures (e.g. pointing coordinates) and outputs the intended output whether spoken (e.g. "This is Hanne's office.") or gestures (e.g. pointing coordinates). Timestamps can include the times a given module commenced and terminated processing and the time a frame was written on the blackboard. The frame semantics also includes representations for two key phenomena in language/vision integration: reference and spatial relations.

Frames can be grouped into three categories: (1) *input*, (2) *output* and (3) *integration*. Input frames are those which come from modules processing perceptual input, output frames are those produced by modules generating system output and integration frames are integrated meaning representations constructed over the course of a dialogue (i.e. all other frames). Here, we shall discuss frames with a focus more on frame semantics than on frame syntax and in fact the actual coding of frames as messages within CHAMELEON has a different syntax (see Appendix A).

3.1 Input frames

An input frame takes the general form:

[MODULE
 INPUT: input
 INTENTION: intention-type
 TIME: timestamp]

where MODULE is the name of the input module producing the frame, INPUT can be at least UTTERANCE or GESTURE, *input* is the utterance or gesture and *intention-type* includes different types of utterances and gestures. An utterance input frame can at least have intention-type (1) query?, (2) instruction! and (3) declarative. An example of an utterance input frame is:

[SPEECH-RECOGNISER
 UTTERANCE: (Point to Hanne's office)
 INTENTION: instruction!
 TIME: timestamp]

A gesture input frame is where *intention-type* can be at least (1) pointing, (2) mark-area, and (3) indicate-direction. An example of a gesture input frame is:

[GESTURE

GESTURE: coordinates (3, 2)
INTENTION: pointing
TIME: timestamp]

3.2 Output frames

An output frame (F-out) takes the general form:

[MODULE
INTENTION: intention-type
OUTPUT: output
TIME: timestamp]

where MODULE is the name of the output module producing the frame, *intention-type* includes different types of utterances and gestures and OUTPUT is at least UTTERANCE or GESTURE. An utterance output frame can at least have intention-type (1) query? (2) instruction!, and (3) declarative. An example utterance output frame is:

[SPEECH-SYNTHESIZER
INTENTION: declarative
UTTERANCE: (This is Hanne's office)
TIME: timestamp]

A gesture output frame can at least have intention-type (1) description (pointing), (2) description (route), (3) description (mark-area), and (4) description (indicate-direction). An example gesture output frame is:

[LASER
INTENTION: description (pointing)
LOCATION: coordinates (5, 2)
TIME: timestamp]

3.3 Integration frames

Integration frames are all those other than input/output frames. An example utterance integration frame is:

[NLP
INTENTION: description (pointing)
LOCATION: office (tenant Hanne) (coordinates (5, 2))
UTTERANCE: (This is Hanne's office)
TIME: timestamp]

Things become even more complex with the occurrence of references and spatial relationships:

[MODULE
INTENTION: intention-type
LOCATION: location
LOCATION: location
LOCATION: location
SPACE-RELATION: beside
REFERENT: person
LOCATION: location
TIME: timestamp]

An example of such an integration frame is:

[DOMAIN-MODEL
INTENTION: query? (who)
LOCATION: office (tenant Hanne) (coordinates (5, 2))
LOCATION: office (tenant Jørgen) (coordinates (4, 2))
LOCATION: office (tenant Børge) (coordinates (3, 1))
SPACE-RELATION: beside
REFERENT: (person Paul-Dalsgaard)
LOCATION: office (tenant Paul-Dalsgaard) (coordinates (4, 1))
TIME: timestamp]

Here we derive all the frames appearing on the blackboard for the example: "Who's in the office beside him?" We have reported complete blackboard histories for the instruction "Point to Hanne's office" and the query "Whose office is this?" + [pointing] (exophoric/deictic reference) in Brøndsted et al. (1998), Mc Kevitt and Dalsgaard (1997), and Mc Kevitt (1997b).

There are input, output and integration frames (F-in, F-out, F-int), input and output gestures (G-in, G-out) and input and output utterances (U-in, U-out). Input modules are SPEECH-RECOGNISER (U-in) and GESTURE (G-in). Output modules are LASER (G-out) and SPEECH-SYNTHESIZER (U-out). Most modules give and take frames to/from the blackboard database and process them (F-int).

We choose to have modules interacting in a completely distributed manner with no single coordinating module. The actual present implementation of CHAMELEON has a dialogue manager which acts as a central coordinator. Although we show the various modules acting in a given sequence here, module processing and frames may not necessarily run in this order. The frames given are placed on the blackboard as they are produced and processed.

3.4 Projective relation "beside"

USER(U-in): Who's in the office beside him?

PROCESSING(1):
SPEECH-RECOGNISER:
(1) wakes up when it detects registering of U-in
(2) maps U-in into F-in

(3) places and registers F-in on blackboard:

FRAME(F-in)(1):
[SPEECH-RECOGNISER
 UTTERANCE: (Who is in the office beside him ?)
 INTENTION: query?
 TIME: timestamp]

PROCESSING(2):
NLP:
(1) wakes up when it detects registering of F-in
(2) maps F-in into F-int
(3) places and registers F-int on blackboard:

FRAME(F-int)(1):
[NLP
 INTENTION: query? (who)
 LOCATION: office (tenant Person) (coordinates (X, Y))
 SPACE-RELATION: beside
 REFERENT: (person him)
 LOCATION: office (tenant Person) (coordinates (X, Y))
 TIME: timestamp]

PROCESSING(3):
DIALOGUE-MANAGER:
(1) wakes up when it detects registering of F-int
(2) reads F-int and sees it's got a reference "him"
(3) searches the blackboard history for the referent "him"
(4) finds the last person mentioned: "Paul Dalsgaard"
(5) produces updated F-int (referent)
(6) places and registers updated F-int on blackboard:

FRAME(F-int)(2):
[DIALOGUE-MANAGER
 INTENTION: query? (who)
 LOCATION: office (tenant Person) (coordinates (X, Y))
 SPACE-RELATION: beside
 REFERENT: (person Paul-Dalsgaard)
 LOCATION: office (tenant Paul-Dalsgaard) (coordinates (X, Y))
 TIME: timestamp]

PROCESSING(4):
DOMAIN-MODEL:
(1) wakes up when it detects registering of F-int
(2) reads F-int and sees its from DIALOGUE-MANAGER
(3) produces updated F-int (space-relation + coordinates)
(4) places and registers filled F-int on blackboard:

FRAME(F-int)(3):
[DOMAIN-MODEL
 INTENTION: query? (who)
 LOCATION: office (tenant Hanne) (coordinates (5, 2))
 LOCATION: office (tenant Jørgen) (coordinates (4, 2))

LOCATION: office (tenant Bøerge) (coordinates (3, 1))
SPACE-RELATION: beside
REFERENT: (person Paul-Dalsgaard)
LOCATION: office (tenant Paul-Dalsgaard) (coordinates (4, 1))
TIME: timestamp]

PROCESSING(5):
NLP:
(1) wakes up when it detects registering of F-int
(2) reads F-int and sees it's from DOMAIN-MODEL
(3) produces updated F-int (intention + utterance)
(4) places and registers updated F-int on blackboard:

FRAME(F-int)(4):
[NLP
INTENTION: declarative (who)
LOCATION: office (tenant Hanne) (coordinates (5, 2))
LOCATION: office (tenant Jørgen) (coordinates (4, 2))
LOCATION: office (tenant Bøerge) (coordinates (3, 1))
SPACE-RELATION: beside
REFERENT: (person Paul-Dalsgaard)
LOCATION: office (tenant Paul-Dalsgaard) (coordinates (4, 1))
UTTERANCE: (Bøerge, Jørgen and Hanne's offices are beside Paul's Dalsgaard's office)
TIME: timestamp]

PROCESSING(6):
LASER:
(1) wakes up when it detects registering of F-int
(2) reads F-int and sees it's from DOMAIN-MODEL
(3) produces F-out (pruning + registering)
(4) places and registers F-out on blackboard:

FRAME(F-out)(1):
[LASER
INTENTION: description (pointing)
LOCATION: coordinates (5, 2)
LOCATION: coordinates (4, 2)
LOCATION: coordinates (3, 1)
SPACE-RELATION: beside
REFERENT: (person Paul-Dalsgaard)
LOCATION: coordinates (4, 1)
TIME: timestamp]

PROCESSING(7):
SPEECH-SYNTHESIZER:
(1) wakes up when it detects registering of F-int
(2) reads F-int and sees it's from NLP
(3) produces F-out (pruning + registering)
places and registers F-out on blackboard:

FRAME(F-out)(2):
[SPEECH-SYNTHESIZER
INTENTION: declarative (who)

UTTERANCE: (Bøerge, Jørgen and Hanne's offices are beside Paul Dalsgaard's office)
TIME: timestamp]

PROCESSING(8):
DIALOGUE-MANAGER:
(1) wakes up when it detects registering of F-out and F-out
(2) reads F-out and F-out and sees they are from
LASER and SPEECH-SYNTHESIZER
(3) dials and fires LASER and SPEECH-SYNTHESIZER
in a rhythmic way (synchronized)
(1) LASER reads its own F-out and fires G-out
(2) SPEECH-SYNTHESIZER reads its own F-out and fires U-out

CHAMELEON(G-out): [points (4 times)]
CHAMELEON(U-out): Bøerge, Jørgen and Hanne's offices are beside Paul Dalsgaard's office.

Note that the above dialogue could also be one where CHAMELEON becomes active and says "There are three offices beside Paul Dalsgaard's, do you mean to the left, in front of or to the right of his office?" This would, of course, involve more complex processing, especially for the dialogue manager.

4 Relation to other work

The representation of the spatial relation "beside" as given in the frame semantics above is similar to what Herskovitz (1996) termed a spatial proposition,

```
(<relation name> <LO> <sequence of ROs>)
```

where `LO` is an object to be localised and `RO` is reference object. In our example above the `LO` is "Paul Dalsgaard" and the `ROs` are the other offices beside his.

Blocher and Stopp (1995) give a detailed computational model for representing and generating spatial relations for SOCCER, a system which automatically generates reports of short soccer games. Their focus is more on generating spatial relations rather than processing them as input queries. Maaβ (1994) looks at the area of route descriptions and how a speaker presents step-by-step relevant route information in a 3D environment with an implementation called MOSES. Specifically addressed is the interaction between the spatial relation and the presentation representation used for natural language descriptions. Again, the focus here is generating spatial relations rather than recognising them. SOCCER and MOSES are part of a general project called VITRA (VIsual TRAnslator) concerning the design and construction of integrated knowledge-based systems for translating visual information into natural language descriptions (Herzog and Wazinski 1994).

The L_0 project (Feldman et al. 1996) focusses on combining not only vision and natural language modelling, but also learning. The task is to build a system that can learn the appropriate fragment of any natural language from sentence-picture pairs. Important lessons have been learned in the subtle semantics of spatial language, especially since L_0 is multilingual (English, Mixtec, German, Bengali, and Japanese) and spatial language is something which changes a lot over languages. The L_0 implementation of spatial language modelling is conducted mainly in the connectionist computational framework.

Situated Artificial Communicators (SFB-360) (Rickheit and Wachsmuth 1996) is a collaborative research project at the University of Bielefeld, Germany which focusses on modelling that which a person performs when with a partner he cooperatively solves a simple assembly task in a given situation. The object chosen is a model airplane (Baufix) to be constructed by a robot from the components of a wooden building kit with instructions from a human. SFB-360 includes equivalents of the modules in CHAMELEON although there is no learning module competitor to Topsy. What SFB-360 gains in size it may loose in integration, i.e. it is not clear yet that all the technology from the subprojects have been fitted together and in particular what exactly the semantic representations passed between the modules are. The DACS process communication system currently used in CHAMELEON is a useful product from SFB-360.

Gandalf is a communicative humanoid which interacts with users in MultiModal dialogue through using and interpreting gestures, facial expressions, body language and spoken dialogue (Thórinson 1997). Gandalf is an application of an architecture called *Ymir* which includes perceptual integration of multimodal events, distributed planning and decision making, layered input analysis and motor-control with human-like characteristics and an inherent knowledge of time. Ymir has a blackboard architecture and includes modules equivalent to those in CHAMELEON. However, there is no vision/image processing module since gesture tracking is done with the use of a data glove and body tracking suit and an eye tracker is used for detecting the user's eye gaze. Also, Ymir has no learning module equivalent to Topsy. Ymir's architecture is even more distributed than CHAMELEON's with many more modules interacting with each other. Also, Ymir's semantic representation is much more distributed with smaller chunks of information than our frames being passed between modules.

AESOPWORLD is an integrated comprehension and generation system for integration of vision, language and motion (Okada 1997). It includes a model of mind consisting of nine domains according to the contents of mental activities and five levels along the process of concept formation. The system simulates the protagonist or fox of an AESOP fable, "the Fox and the Grapes", and his mental and physical behaviour are shown by graphic displays, a voice generator, and a music generator which expresses his emotional states. AESOPWORLD has an agent-based distributed architecture and also uses frames as semantic representations. It has many modules in common with CHAMELEON although again there is no vision input to AESOPWORLD which uses computer graphics to depict scenes. AESOPWORLD has an extensive planning module but conducts more traditional planning than CHAMELEON's Topsy.

The INTERACT project (Waibel et al. 1996) involves developing MultiModal Human Computer Interfaces including the modalities of speech, gesture and pointing, eye-gaze, lip motion and facial expression, handwriting, face recognition and tracking, and sound localisation. The main concern is with improving recognition accuracies of modality specific component processors as well as developing optimal combinations of multiple input signals to deduce user intent more reliably in cross-modal speech-acts. INTERACT also uses a frame representation for integrated semantics from gesture and speech and partial hypotheses are developed in terms of partially filled frames. The output of the interpreter is obtained by unifying the information contained in the partial frames. Although Waibel et al. present good work on multimodal interfaces it is not clear that they have developed an integrated platform which can be used for developing multimodal applications.

5 Conclusion and future work

We have described the architecture and implementation of CHAMELEON: an open, distributed architecture with ten modules glued into a single platform using the DACS communication system. Also described is the IntelliMedia WorkBench application, a software and physical platform where a user can ask for information about things on a physical table and, in particular, the Campus Information System domain. Next, we discussed the frame semantics representation of CHAMELEON and how the query, "Who's in the office beside him?" is processed through the semantics with all associated frames and module interactions. More details on CHAMELEON and the IntelliMedia WorkBench can be found in Brøndsted et al. (1998).

There are a number of avenues for future work with CHAMELEON. The frame semantics handling of "beside" has yet to be implemented and the next step is to move onto modelling other projective spatial relations. Also, presently CHAMELEON provides route descriptions through laser pointing but also more detailed verbal descriptions could be given hand-in-hand with those drawn by the laser, mentioning "left", "right" and other turns for routes. It is hoped that more complex decision taking can be introduced to operate over semantic representations in the dialogue manager or blackboard using, for example, the HUGIN software tool (Jensen (F.) 1996) based on Bayesian Networks (Jensen (F.V.) 1996). The gesture module will be augmented so that it can handle gestures other than pointing. Topsy will be asked to do more complex learning and processing of input/output from frames. The microphone array has to be integrated into CHAMELEON and set to work.

Intelligent MultiMedia will be important in the future of international computing and media development and IntelliMedia 2000+ at Aalborg University, Denmark brings together the necessary ingredients from research, teaching and links to industry to enable its successful implementation. Our CHAMELEON platform and IntelliMedia WorkBench application are ideal for testing integrated processing of language and vision for the future of SuperinformationhighwayS.

6 Acknowledgements

This opportunity is taken to acknowledge support from the Faculty of Science and Technology, Aalborg University, Denmark and Paul Mc Kevitt would also like to acknowledge the British Engineering and Physical Sciences Research Council (EPSRC) for their generous funded support under grant B/94/AF/1833 for the Integration of Natural Language, Speech and Vision Processing (Advanced Fellow) and LIMSI-CNRS, Orsay, France where he was a Visiting Professor whilst completing this paper. Annelies Braffort, Tom Brøndsted, Paul Dalsgaard, Rachid Gherbi, Lars Bo Larsen, Michael Manthey, Thomas B. Moeslund, and Kristian G. Olesen are acknowledged for useful discussions.

Appendix A

Blackboard in practice

Here we show the complete blackboard (with all frames) as produced exactly by CHAMELEON for the example dialogue given in Section 2.

```
Received: nlp(intention(instruction(pointing)),location(person(tb),type(office)),
time(889524794))
which is passed on to dialog_manager
Received: dialog_manager(output(laser(point(coordinates(249,623))),
speech_synthesizer(utterance("This is Toms office"))))
Calling laser: laser(point(coordinates(249,623)))
Calling speech_synthesizer: speech_synthesizer(utterance("This is Toms office"))


Received: nlp(intention(instruction(pointing)),location(person(tbm),type(office)),
time(889524818))
which is passed on to dialog_manager
Received: dialog_manager(output(laser(point(coordinates(278,623))),
speech_synthesizer(utterance("This is Thomass office"))))
Calling laser: laser(point(coordinates(278,623)))
Calling speech_synthesizer: speech_synthesizer(utterance("This is Thomass office"))


Received: nlp(intention(query(where)),location(place(a2_221)),
time(889524831))
which is passed on to dialog_manager
Received: dialog_manager(output(laser(point(coordinates(132,500))),
speech_synthesizer(utterance("computer room is here"))))
Calling laser: laser(point(coordinates(132,500)))
Calling speech_synthesizer: speech_synthesizer(utterance("computer room is here"))


Received: nlp(intention(query(who)),location(this($Deixis),type(office)),
time(889524864))
which is passed on to dialog_manager
Received: dialog_manager(output(laser(point(coordinates(658,546))),
speech_synthesizer(utterance("This is not an office, this is instrument repair"))))
Calling laser: laser(point(coordinates(658,546)))
Calling speech_synthesizer:
speech_synthesizer(utterance("This is not an office, this is instrument repair"))


Received: nlp(intention(query(who)),location(this($Deixis),type(office)),
time(889524885))
```

```
which is passed on to dialog_manager
Received: dialog_manager(output(laser(point(coordinates(223,568))),
speech_synthesizer(utterance("This is Pauls office"))))
Calling laser: laser(point(coordinates(223,568)))
Calling speech_synthesizer: speech_synthesizer(utterance("This is Pauls office"))


Received: nlp(intention(instruction(show_route)),source(location(person(lbl),
type(office))),
destination(location(person(hg),type(office))),time(889524919))
which is passed on to dialog_manager
Received: dialog_manager(output(laser(route(coordinates(278,585,278,603,249,
603,220,603,197,603,197,623))),
speech_synthesizer(utterance("This is the route from Lars Bos office to Hannes office"))))
Calling laser:
laser(route(coordinates(278,585,278,603,249,603,220,603,197,603,197,623)))
Calling speech_synthesizer:
speech_synthesizer(utterance("This is the route from Lars Bos office to Hannes office"))


Received: nlp(intention(instruction(show_route)),source(location(person(pmck),
type(office))),destination(location(place(a2_105))),time(889524942))
which is passed on to dialog_manager
Received: dialog_manager(output(laser(route(coordinates(174,453,153,453,153,
481,153,500,153,510,153,540,153,569,153,599,153,603,184,603,197,603,220,603,
249,603,278,603,307,603,330,603,330,655,354,655,911,655,884,655,884,603,810,
603,759,603,717,603,717,570,696,570))),
speech_synthesizer(utterance("This is the route from Pauls office to instrument repair"))))
Calling laser: laser(route(coordinates(174,453,153,453,153,481,153,500,153,
510,153,540,153,569,153,599,153,603,184,603,197,603,220,603,249,603,278,603,
307,603,330,603,330,655,354,655,911,655,884,655,884,603,810,603,759,603,717,
603,717,570,696,570)))
Calling speech_synthesizer:
speech_synthesizer(utterance("This is the route from Pauls office to instrument repair"))


Received: nlp(intention(instruction(pointing)),location(person(pd),type(office)),
time(889524958))
which is passed on to dialog_manager
Received: dialog_manager(output(laser(point(coordinates(220,585))),
speech_synthesizer(utterance("This is Pauls office"))))
```

References

Brøndsted (1998) *nlparser*. WWW: http://www.kom.auc.dk/~tb/nlparser.

Brøndsted, T., P. Dalsgaard, L.B. Larsen, M. Manthey, P. Mc Kevitt, T.B. Moeslund, K.G. Olesen (1998) *A platform for developing Intelligent MultiMedia applications*. Technical Report R-98-1004, Center for PersonKommunikation (CPK), Institute for Electronic Systems (IES), Aalborg University, Denmark, May.

Christensen, Heidi, Børge Lindberg and Pall Steingrimsson (1998) *Functional specification of*

the CPK Spoken LANGuage recognition research system (SLANG). Center for PersonKommunikation, Aalborg University, Denmark, March.

Denis, M. and M. Carfantan (Eds.) (1993) *Images et langages: multimodalité et modelisation cognitive.* Actes du Colloque Interdisciplinaire du Comité National de la Recherche Scientifique, Salle des Conférences, Siège du CNRS, Paris, April.

Feldman, Jerome, George Lakoff, David Bailey, Srini Narayanan, Terry Regier and Andreas Stolcke (1996) L_0 – the first five years of an automated language acquisition project. In *Artificial Intelligence Review*, 8, 175-187.

Fink, Gernot A., Nils Jungclaus, Franz Kummert, Helge Ritter and Gerhard Sagerer (1996) A distributed system for integrated speech and image understanding. In *Proceedings of the International Symposium on Artificial Intelligence*, Rogelio Soto (Ed.), 117-126. Cancun, Mexico.

Herskovitz (1989) *Language and spatial cognition: an interdisciplinary study of the prepositions in English.* Cambridge, England: Cambridge University Press.

Herzog, Gerd and Peter Wazinski (1994) VIsual TRAnslator: linking perceptions and natural language descriptions. In *Artificial Intelligence Review*, 8, 175-187.

Infovox (1994) *INFOVOX: Text-to-speech converter user's manual (version 3.4).* Solna, Sweden: Telia Promotor Infovox AB.

Jensen, Finn V. (1996) *An introduction to Bayesian Networks.* London, England: UCL Press.

Jensen, Frank (1996) Bayesian belief network technology and the HUGIN system. In *Proceedings of UNICOM seminar on Intelligent Data Management*, Alex Gammerman (Ed.), 240-248. Chelsea Village, London, England, April.

Kosslyn, S.M. and J.R. Pomerantz (1977) Imagery, propositions and the form of internal representations. In *Cognitive Psychology*, 9, 52-76.

Leth-Espensen, P. and B. Lindberg (1996) Separation of speech signals using eigenfiltering in a dual beamforming system. In *Proc. IEEE Nordic Signal Processing Symposium (NORSIG)*, Espoo, Finland, September, 235-238.

Maaβ, Wolfgang (1994) From vision to multimodal communication: incremental route descriptions. In *Artificial Intelligence Review*, 8, 159-174.

Maaβ, Wolfgang (Ed.) (1996) *Proceedings of the ECAI-96 Workshop on the representations and processes between vision and natural language.* Twelfth European Conference on Artificial Intelligence (ECAI-96), Budapest, Hungary.

Manthey, Michael J. (1998) The Phase Web Paradigm. In *International Journal of General Systems, special issue on General Physical Systems Theories*, K. Bowden (Ed.). in press.

Mc Kevitt, Paul (1994) Visions for language. In *Proceedings of the Workshop on Integration of Natural Language and Vision processing*, Twelfth American National Conference on Artificial Intelligence (AAAI-94), Seattle, Washington, USA, August, 47-57.

Mc Kevitt, Paul (Ed.) (1995/1996) *Integration of Natural Language and Vision Processing (Vols. I-IV).* Dordrecht, The Netherlands: Kluwer-Academic Publishers.

Mc Kevitt, Paul (1997a) SuperinformationhighwayS. In *"Sprog og Multimedier" (Speech and Multimedia)*, Tom Brøndsted and Inger Lytje (Eds.), 166-183, April 1997. Aalborg, Denmark: Aalborg Universitetsforlag (Aalborg University Press).

Mc Kevitt, Paul (1997b) IntelliMedia TourGuide: understanding reference at the language/vision interface. In *Proceedings of the European Science Foundation (ESF) Network on Converging Computing Methodologies in Astronomy (CCMA)*, Final Conference on Advanced Techniques and Methods for Astronomical Information Handling, M.C. Maccarone, F. Murtagh, M. Kurtz and A. Bijaoui (Eds.), 93-103. Sonthofen, Germany, September, Observatoire de la

Cote d' Azur - B.P. 4229 06304 Cedex 4, France. (also published on Web; see http://newb6.u-strasbg.fr/~ccma).

Mc Kevitt, Paul and Paul Dalsgaard (1997) A frame semantics for an IntelliMedia TourGuide. In *Proceedings of the Eighth Ireland Conference on Artificial Intelligence (AI-97), Volume 1*, 104-111. University of Uster, Magee, Derry, Northern Ireland, September.

Minsky, Marvin (1975) A framework for representing knowledge. In *The Psychology of Computer Vision*, P.H. Winston (Ed.), 211-217. New York: McGraw-Hill.

Nielsen, Claus, Jesper Jensen, Ove Andersen, and Egon Hansen (1997) *Speech synthesis based on diphone concatenation.* Technical Report, No. CPK971120-JJe (in confidence), Center for PersonKommunikation, Aalborg University, Denmark.

Okada, Naoyuki (1997) Integrating vision, motion and language through mind. In *Proceedings of the Eighth Ireland Conference on Artificial Intelligence (AI-97), Volume 1*, 7-16. University of Uster, Magee, Derry, Northern Ireland, September.

Olivier, Patrick Luke (Ed.) (1995) *Working notes of the workshop on Representation and processing of spatial expressions.* Fourteenth International Joint Conference on Artificial Intelligence (IJCAI-95), Montreal, Canada, August.

Olivier, Patrick Luke (Ed.) (1996) *Proceedings of the ECAI-96 workshop on the representations and processing of spatial expressions.* Twelfth European Conference on Artificial Intelligence (ECAI-96), Budapest, Hungary.

Olivier, Patrick Luke (Ed.) (1997) *Proceedings of the AAAI-97 workshop on language and space.* Fourteenth American National Conference on Artificial Intelligence (AAAI-97), Rhode Island, New Jersey.

Pentland, Alex (Ed.) (1993) *Looking at people: recognition and interpretation of human action.* IJCAI-93 Workshop (W28) at The 13th International Conference on Artificial Intelligence (IJCAI-93), Chambéry, France, EU, August.

Power, Kevin, Caroline Matheson, Dave Ollason and Rachel Morton (1997) *The grapHvite book (version 1.0).* Cambridge, England: Entropic Cambridge Research Laboratory Ltd..

Pylyshyn, Zenon (1973) What the mind's eye tells the mind's brain: a critique of mental imagery. In *Psychological Bulletin*, 80, 1-24.

Retz-Schmidt, Gudala (1988) Various views on spatial prepositions. In *AI Magazine*, 9: 95-105.

Rickheit, Gert and Ipke Wachsmuth (1996) Collaborative Research Centre "Situated Artificial Communicators" at the University of Bielefeld, Germany. In *Integration of Natural Language and Vision Processing, Volume IV, Recent Advances*, Mc Kevitt, Paul (ed.), 11-16. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Thórinsson, Kris R. (1997) Layered action control in communicative humanoids. In *Proceedings of Computer Graphics Europe '97*, June 5-7, Geneva, Switzerland.

Waibel, Alex, Minh Tue Vo, Paul Duchnowski and Stefan Manke (1996) Multimodal interfaces. In *Integration of Natural Language and Vision Processing, Volume IV, Recent Advances*, Mc Kevitt, Paul (Ed.), 145-165. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Zelinsky-Wibbelt (Ed.), Cornelia (1993) *The semantics of prepositions: from mental Processing to natural language processing (NLP 3).* Berlin, Germany: Mouton de Gruyter.